

(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION
EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété
Intellectuelle
Bureau international



(43) Date de la publication internationale
29 mars 2001 (29.03.2001)

PCT

(10) Numéro de publication internationale
WO 01/22279 A1

(51) Classification internationale des brevets⁷: G06F 17/30

(21) Numéro de la demande internationale:

PCT/FR00/02640

(22) Date de dépôt international:

22 septembre 2000 (22.09.2000)

(25) Langue de dépôt:

français

(26) Langue de publication:

français

(30) Données relatives à la priorité:

99/11973

24 septembre 1999 (24.09.1999)

FR

(71) Déposant (pour tous les États désignés sauf US):
FRANCE TELECOM [FR/FR]; 6, place d'Alleray,
F-75015 Paris (FR).

(72) Inventeurs; et

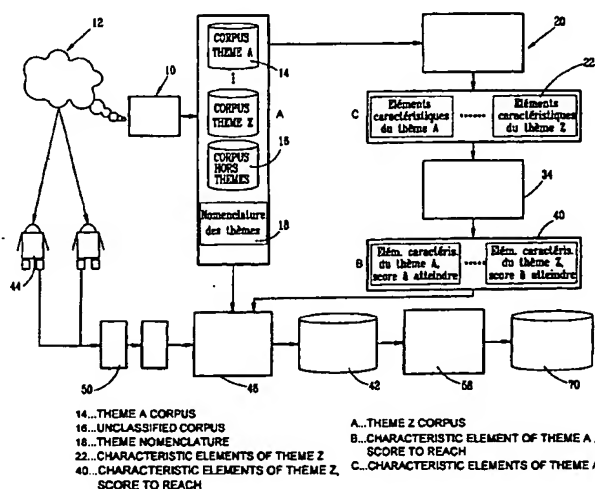
(75) Inventeurs/Déposants (pour US seulement): **BIET-
TRON**, Laurent [FR/FR]; Kerauzern, F-22300 Ploubezre
(FR). **PALLU**, Frédéric [FR/FR]; Keravel, F-22560
Trebeurden (FR). **TRICOT**, Sylvie [FR/FR]; 14, Hent
Lann, F-22300 Tredrez (FR).

(74) Mandataires: **JACOBSON**, Claude etc.; Cabinet Lavoix,
2, place d'Estienne d'Orves, F-75441 Paris Cedex 09 (FR).

[Suite sur la page suivante]

(54) Title: METHOD FOR THEMATIC CLASSIFICATION OF DOCUMENTS, THEMATIC CLASSIFICATION MODULE
AND SEARCH ENGINE INCORPORATING SUCH A MODULE

(54) Titre: PROCEDE DE CLASSIFICATION THEMATIQUE DE DOCUMENTS, MODULE DE CLASSIFICATION THEMA-
TIQUE ET MOTEUR DE RECHERCHE INCORPORANT UN TEL MODULE



(57) Abstract: The invention concerns a method for thematic classification of documents, in particular for constituting or updating thematic databases (42) for a search engine, comprising steps which consist in: selecting documents representing each theme; identifying in the selected documents elements characteristic of each theme; assigning to each identified element a coefficient (R) representing the relevance of said element relative to the corresponding theme; and, for each document (50) to be classified, identifying said elements characteristic of each theme it contains, and for each theme which corresponds to them, computing from the coefficient assigned to said elements, the value of a characteristic representative the relevance to the theme for said document (50) to determine whether said document is related or not to said theme.

(57) Abrégé: Ce procédé de classification thématique de documents, notamment pour la constitution ou la mise à jour de bases de données thématiques (42) pour moteur de recherche, comprend les étapes de sélection de documents représentatifs de chaque thème, identification, dans les documents sélectionnés, des éléments caractéristiques de

[Suite sur la page suivante]

WO 01/22279 A1



(81) États désignés (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), brevet OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Publiée:

— Avec rapport de recherche internationale.

(84) États désignés (*régional*): brevet ARIPO (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), brevet eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), brevet européen

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

chaque thème, affectation, à chaque élément identifié, d'un coefficient (R) représentatif de la pertinence de cet élément vis à vis du thème correspondant et, pour chaque document (50) à classer, identification desdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, calcul, à partir du coefficient affecté à ces éléments, de la valeur d'une caractéristique représentative de la pertinence du thème pour ce document (50) pour décider si ce document porte ou non sur ce thème.

Procédé de classification thématique de documents, module de
classification thématique et moteur de recherche incorporant
un tel module

La présente invention se rapporte à un procédé de classification thématique de documents, destiné, en particulier, à la constitution ou la mise à jour de bases de données thématiques, en particulier pour moteur de recherche.

Elle se rapporte également à un module de classification thématique de documents et à un moteur de recherche équipé d'un tel module de classification thématique.

On connaît, à ce jour, principalement deux outils informatiques permettant de rechercher des documents sur un réseau informatique, comme par exemple, le réseau Internet.

Ces outils sont le moteur de recherche et le guide.

Un moteur de recherche est un outil permettant d'extraire d'une information, principalement textuelle, les mots ou termes qui la représentent le mieux et de les stocker dans des bases de données, également connues sous l'appellation "base d'index".

De telles bases d'index sont généralement mises à jour relativement fréquemment.

En réponse à une requête formulée par un utilisateur, ce même outil parcourt les bases d'index afin d'identifier les termes les plus pertinents par rapport à ceux de la requête, puis de trier les informations à fournir en retour.

L'autre technique de recherche de documents sur un réseau informatique consiste à utiliser un guide. Cet outil propose des recherches par catégories, les pages de documents étant classées manuellement par des documentalistes.

Ces types d'outil présentent un certain nombre d'inconvénients.

Tout d'abord, les moteurs de recherche ne proposent pas de classement de pages de document par catégories. En effet, les pages fournies en réponse à une requête ne sont pas typées. Ainsi, des requêtes ambiguës peuvent

donner lieu à des réponses très diverses, ressenties comme du bruit par l'utilisateur.

Les guides, au contraire, permettent de fournir à un utilisateur des réponses typées, c'est à dire portant sur
5 le ou les mêmes thèmes que la requête.

Cependant, le classement manuel des pages de document implique de forts coûts de création et de mise à jour et ne permet l'indexation que d'un nombre limité de pages. Par conséquent, certaines requêtes n'obtiennent
10 pas de réponse.

Le but de l'invention est de palier les inconvénients des moteurs de recherche et des guides.

Elle a donc pour objet un procédé de classification thématique de documents, notamment pour la constitution
15 ou la mise à jour de bases de données thématiques pour moteur de recherche, caractérisé en ce qu'il comporte les étapes suivantes :

- on sélectionne un échantillon de documents représentatifs de chaque thème ;
- 20 - on identifie, dans les documents sélectionnés, des éléments caractéristiques de chaque thème ;
- on affecte, à chaque élément identifié, un coefficient représentatif de la pertinence de cet élément vis à vis du thème correspondant ;
- 25 - pour chaque document à classifier, on identifie lesdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, on calcule, à partir du coefficient affecté à ces éléments, la valeur d'une caractéristique représentative de la
30 pertinence du thème pour ce document, pour décider si ce document porte ou non sur ce thème, lesdites étapes d'identification et de calcul étant réalisées automatiquement pour chaque document récupéré sur un réseau informatique ;
- 35 - on classe les documents récupérés en fonction des thèmes qui y sont abordés ; et

- l'on stocke les documents classés par thèmes dans des bases de données interrogeables à partir de thèmes contenus dans une requête ;

et en ce que l'étape d'affectation dudit coefficient à chaque élément identifié comprend les étapes suivantes, pour chaque thème :

- calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème ;

- calcul de la fréquence de l'élément dans les documents sélectionnés ne portant pas sur ce thème ; et

- calcul du rapport entre les fréquences calculées.

On classe ainsi les documents récupérés sur un réseau informatique en fonction des thèmes qui y sont abordés et ce, de façon automatique.

Le procédé de classification selon l'invention peut en outre comporter une ou plusieurs des caractéristiques suivantes, prises isolément ou selon toutes les combinaisons techniquement possibles :

- il comporte en outre une étape de tri des thèmes selon une arborescence de thèmes et par ordre décroissant des coefficients ;

- l'étape de calcul de la caractéristique représentative de la pertinence du thème d'un document à classer comprend les étapes suivantes pour chaque thème :

. on lit la valeur du rapport desdites fréquences de chaque élément représentatif du thème extrait du document,

. on multiplie les valeurs lues, et

. on affecte le résultat de cette multiplication à la valeur de ladite caractéristique ;

- l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil ;

- la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

score - seuil_{thème} = (R_{moy})nthème

dans laquelle :

score - seuil_{thème} désigne la valeur de seuil

R_{moy} représente la valeur moyenne des rapports de

5 fréquences R des éléments du thème et,

nthème désigne un nombre prédéterminé ;

- selon une variante, la valeur de seuil est réglée
manuellement ;

- les étapes d'identification des éléments
10 caractéristiques de chaque thème contenu dans un document
sont réalisées au moyen d'une table de hachage ; et

- on calcule, pour chaque élément de vocabulaire
d'une requête formulée par un utilisateur, des
coefficients caractéristiques de l'élément par rapport à
15 chaque thème connu et l'on associe à chaque élément les
coefficients et les thèmes correspondants, de sorte que
lesdits coefficients atteignent une valeur minimale.

Lors de la recherche des entrées d'index, c'est à
dire au cours de la recherche des documents
20 correspondants à la requête, il est ainsi possible
d'accéder directement aux thèmes liés à chaque élément et
aux coefficients correspondants que l'on combine par
multiplication afin de déterminer un classement des
thèmes liés à la requête entière.

25 L'invention a également pour objet un module de
classification thématique de documents, notamment pour
moteur de recherche, caractérisé en ce qu'il comporte une
unité centrale de traitement comprenant des moyens de
comparaison d'éléments extraits de chaque document avec
30 des éléments caractéristiques de différents thèmes,
affectés chacun d'un coefficient représentatif de la
pertinence de cet élément pour un thème correspondant, et
des moyens de calcul de la valeur d'au moins une
caractéristique représentative de la pertinence d'un
35 thème pour ce document, à partir des coefficients desdits
éléments caractéristiques qu'il contient, pour décider si
ce document porte ou non sur ce thème, ladite unité
centrale étant raccordée à des moyens de stockage de

documents classés par thèmes, interrogeables à partir de thèmes contenus dans une requête, et en ce qu'il comporte des moyens de calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème, des
5 moyens de calcul de la fréquence de l'élément dans les documents sélectionnés ne portant pas sur ce thème, et des moyens de calcul du rapport entre les fréquences calculées.

Un autre objet de l'invention est un moteur de
10 recherche de documents sur un réseau informatique, comprenant un module d'indexation pour la création et la mise à jour de bases de données thématiques, à partir de documents récupérés sur le réseau informatique, et un module d'interrogation des bases de données adaptées pour
15 fournir des références de documents correspondant à une requête reçue en entrée, caractérisé en ce qu'il comporte en outre un module de classification thématique tel que définit ci-dessus, associé au module d'indexation.

D'autres caractéristiques et avantages ressortiront
20 de la description suivante, donnée uniquement à titre d'exemple, et faite en référence aux dessins annexés sur lesquels :

- la Fig. 1 est un organigramme montrant les principales phases de fonctionnement d'un module de
25 classification thématique de documents selon l'invention, pour moteur de recherche ;

- la Fig. 2 est un organigramme illustrant la
 . méthode de calcul des éléments caractéristiques de thèmes ; et

30 - la Fig. 3 est un organigramme montrant la méthode de calcul des thèmes d'un document.

Sur la Fig. 1, on a représenté les principales phases du procédé de classification thématique de documents selon l'invention.

35 Il est destiné à permettre le classement de documents récupérés sur un réseau informatique, en fonction de thèmes qui y sont abordés. Par exemple, il peut être mis en oeuvre au sein d'un moteur de recherche.

Dans ce cas, il intervient dès le processus d'indexation, mais également au cours du traitement d'une requête formulée par un utilisateur, pour permettre de déterminer tous les thèmes abordés dans cette requête.

5 On conçoit toutefois que d'autres applications peuvent être envisagées. Par exemple, ce procédé peut être mis en oeuvre au niveau d'un point d'accès d'un réseau de postes utilisateurs à un réseau Internet, afin de déterminer la nature des pages Web récupérées par les
10 utilisateurs et interdire ou autoriser, par filtrage des requêtes, certains thèmes, par exemple, contraires à l'ordre public et aux bonnes moeurs, ou encore calculer des statistiques sur les centres d'intérêt des utilisateurs.

15 Pour procéder à cette classification, le procédé comporte deux phases distinctes, à savoir une première phase préalable d'acquisition du vocabulaire thématique de corpus de documents et d'affectation, à chaque mot du vocabulaire, d'une valeur de seuil à partir de laquelle
20 on décide qu'un document, contenant ce mot, porte sur le thème correspondant, ainsi qu'une deuxième phase de classification proprement dite, au cours de laquelle un document récupéré sur le réseau est automatiquement classifié en fonction des éléments caractéristiques qu'il
25 contient.

Par exemple cette deuxième phase intervient périodiquement, seuls des documents nouvellement créés ou modifiés étant classifiés.

La description de la première phase d'acquisition du
30 vocabulaire thématique va maintenant être en référence aux Figs. 1 à 3.

Comme on le voit sur la Fig. 1, cette phase débute par une étape 10 de sélection manuelle, à partir d'un ensemble 12 d'échantillons (ou corpus) de documents
35 représentatifs de chacun des thèmes A à Z utilisés pour classer les documents au cours de la deuxième phase.

Ainsi, à l'issu de cette étape 10 de sélection manuelle, on dispose d'un ensemble de corpus de

documents, tels que 14, portant chacun sur un thème (thème A,... thème Z). Bien entendu l'étape de sélection peut également être effectuée par tout moyen autre que manuel.

5 Au cours de cette étape 10 de sélection, on crée également un corpus 16 de documents ne portant sur aucun des thèmes A à Z et on définit une nomenclature 18 des thèmes A à Z, c'est à dire la liste de ces thèmes associés à des sous-thèmes s'y rapportant.

10 Lors de l'étape 20 suivante, ces éléments sont présentés en entrée d'un module de classification thématique en vue d'extraire de chaque document les éléments caractéristiques de chaque thème et de les affecter chacun d'un coefficient représentatif de leur
15 pertinence vis à vis d'un thème correspondant.

Par exemple ce module de classification thématique se présente sous la forme d'un module spécifique d'un moteur de recherche, associé à un module d'indexation réalisant la création ou la mise à jour des bases de
20 données thématiques.

Il peut également être agencé sous la forme d'un module spécifique prévu au niveau d'un point d'accès à un réseau informatique, en particulier à un réseau Internet.

Ce module comprend les moyens logiciels appropriés
25 pour réaliser l'extraction des éléments caractéristiques de chaque thème et pour les affecter d'un coefficient représentatif de leur pertinence vis à vis de différents thèmes, comme cela va être décrit en détail par la suite.

Au cours de cette étape 20, le module de
30 classification extrait, de chaque document sélectionné, les éléments caractéristiques de chaque thème.

Cette extraction s'effectue en utilisant un outil informatique de type classique. Il ne sera donc pas décrit par la suite.

35 On dispose à l'issue de cette étape 20, de listes d'éléments caractéristiques des thèmes A à Z, telles que 22.

En référence à la Fig. 2, cette procédure d'identification du vocabulaire caractéristique de chaque thème s'effectue successivement pour chaque élément extrait des documents de chacun des corpus 14 et 16.

5 Au cours d'une première étape 24, on vide un tableau regroupant l'ensemble des thèmes candidats, c'est à dire les thèmes susceptibles de correspondre à l'élément extrait.

10 Lors de l'étape 26 suivante, on procède, pour chaque thème, à un calcul d'un coefficient R représentatif de la pertinence de cet élément vis à vis de ce thème.

 Pour procéder à ce calcul, on calcule tout d'abord la fréquence p de l'élément dans les documents portant sur ce thème, ainsi que la fréquence q de cet élément
15 dans les documents ne portant pas sur ce thème.

 On procède ensuite au calcul du coefficient R, constitué par le rapport entre ces fréquences p et q.

 Lors de l'étape 28 suivante, on vérifie si les caractéristiques p, q et R se situent à l'intérieur de
20 limites prédéterminées.

 Si tel n'est pas le cas, on procède au traitement de l'élément suivant.

 Si tel est le cas, on ajoute le thème dans le tableau des thèmes candidats avec un score égal au
25 coefficient R (étape 30).

 S'il reste des éléments à traiter (étape 32), la procédure retourne à l'étape 24 précédente.

 Dans le cas contraire, cette procédure s'achève.

 On notera que, de préférence, après remplissage du
30 tableau des thèmes candidats, celui-ci est trié par ordre décroissant des scores R. On notera également que pour tout thème candidat, jusqu'à un nombre maximum voulu, on ajoute un nouvel élément récupéré dans la liste des éléments caractéristiques de ce thème, en se limitant à
35 un nombre maximum voulu des n meilleurs éléments par thème choisi en fonction de leur score R.

 En se référant à nouveau à la Fig. 1, lors de l'étape 34 suivante, le module de classification

thématique procède à un calcul automatique, au moyen d'un algorithme approprié, d'une valeur de seuil correspondant à un seuil minimum à atteindre pour déterminer si un document comprenant un élément caractéristique d'un thème
5 porte ou non sur ce thème.

Pour procéder à ce calcul, le module de classification procède tout d'abord à un calcul de la valeur moyenne R_{moy} des rapports R des éléments caractéristiques de chaque thème (étape 36).

10 Il procède ensuite au calcul de la valeur de seuil score - seuil_{thème}, selon la relation suivante :

$$\text{score - seuil}_{\text{thème}} = (R_{\text{moy}})n_{\text{thème}}$$

dans laquelle $n_{\text{thème}}$ désigne un nombre prédéterminé choisi par exemple égal à 5 pour la plupart des thèmes.

15 On voit alors sur la Fig. 1, qu'à l'issu de ce calcul automatique des scores à atteindre, on dispose de listes, telles que 40, d'éléments caractéristiques de chaque thème A à Z, affectés chacun d'un score à atteindre, c'est à dire d'une valeur de seuil à partir de
20 laquelle on considère qu'un document porte sur ce thème.

Après cette phase d'acquisition du vocabulaire thématique, réalisée à partir de corpus de documents représentatifs de thèmes, la deuxième phase de classification thématique proprement dite peut être
25 effectuée, dans le but de constituer des bases de données thématiques, désignées par la référence numérique générale 42, à partir de documents collectés automatiquement sur le réseau informatique par des robots, tels que 44.

30 Ces documents sont présentés en entrée du module de classification thématique, qui reçoit également une indication de la nomenclature 18 des thèmes, ainsi que les éléments disponibles à l'issu de l'étape 34 mentionnée précédemment. Ce module procède à un calcul
35 automatique des thèmes sur lesquels porte le document (étape 46).

Pour ce faire, il comporte tous les moyens logiciels appropriés pour réaliser les opérations mentionnées ci-dessous.

En référence à la Fig. 3, au cours d'une première
5 étape 48 de cette procédure, le module d'indexation extrait de chaque document 50 récupéré par les robots 44, les éléments caractéristiques de thèmes qu'il contient.

Cette étape s'effectue, par exemple, en utilisant une table de hachage, pour rechercher rapidement dans les
10 listes d'éléments caractéristiques les éléments contenus dans chaque document.

Après extraction de ces éléments on identifie, parmi ceux-ci, les éléments caractéristiques de thèmes contenus dans les listes 40.

15 Pour chaque élément identifié, le module de classification procède ensuite à un calcul d'une valeur caractéristique représentative de la pertinence de chaque thème pour ce document, à partir du coefficient affecté à cet élément.

20 Pour ce faire, lors de l'étape 52 suivante, une variable "score-thème", représentative du score du document dans un thème donné est positionnée à 1, et ce pour tous les thèmes.

Ensuite, pour tout élément du document, et pour
25 chaque thème de l'arborescence des thèmes, si l'élément se situe parmi la liste des éléments caractéristiques du thème, on lit le score R, c'est à dire la valeur du rapport des fréquences pour chaque élément et on multiplie les valeurs lues du score R pour chacun de ces
30 éléments.

Le résultat de cette multiplication est ensuite affecté à la valeur de la caractéristique score - thème (étape 54).

On décide alors que les thèmes reconnus dans le
35 document 50 sont ceux dont la caractéristique score - thème atteint ou dépasse le score à atteindre pour ces thèmes (étape 56).

On dispose alors, à l'issu de cette procédure, de l'ensemble 57 des thèmes sur le ou lesquels porte le document 50 récupéré.

On conçoit donc que cette procédure de calcul automatique des thèmes des documents récupérés par les robots 44 permet au module d'indexation d'un moteur de recherche de classer ces documents en fonction des thèmes abordés et de constituer les bases 42 de données thématiques.

Une telle procédure de calcul automatique de thème de documents peut également être utilisée pour déterminer les thèmes abordés dans une requête formulée par un utilisateur.

Pour ce faire, à partir de cette requête, pour chacun des éléments du vocabulaire d'interrogation utilisés dans la requête, on calcule les coefficients caractéristiques de cet élément par rapport à chacun des thèmes connus et l'on associe à chacun de ces éléments les coefficients et thèmes de telle manière que les coefficients atteignent une valeur minimale.

Lors de la recherche des entrées d'index correspondant aux éléments d'une requête, c'est à dire pour le calcul des résultats, on accède ainsi directement au thème lié aux éléments ainsi qu'à leur coefficient, que l'on combine par multiplication, selon la même procédure que celle décrite plus haut, afin de déterminer un classement des thèmes liés à la requête entière.

On conçoit donc que cette procédure permet de proposer à un utilisateur de préciser sa requête, par exemple, lorsque celle-ci est formulée de façon vague.

On conçoit également que cette procédure, qui permet d'identifier les thèmes contenus dans une requête, rend possible d'effectuer une surveillance des requêtes utilisateurs afin d'établir des calculs statistiques permettant de définir des profils d'utilisateurs en fonction des requêtes.

On saisira alors que l'invention qui vient d'être décrite peut être utilisée pour la recherche de thèmes

contenus dans des pages récupérées sur un réseau informatique, pour la détermination de thèmes contenus dans une requête formulée par un utilisateur et, à partir de cette détermination, pour le filtrage des requêtes et également des pages récupérées, afin d'interdire la formulation de requête ou la récupération de pages portant sur des thèmes prédéterminés interdits, et pour l'élaboration des profils d'utilisateurs.

On notera cependant que dans le cas de la détermination des thèmes contenus dans une requête, cette dernière est considérée comme constituant un document présenté en entrée du module de classification thématique selon l'invention.

L'invention n'est pas limitée au mode de réalisation envisagée.

En effet, il est également possible, en variante, de régler manuellement la valeur de seuil à partir de laquelle on décide qu'un document porte ou non sur un thème donné.

REVENDEICATIONS

1. Procédé de classification thématique de documents, notamment pour la constitution ou la mise à jour de bases de données thématiques pour moteur de recherche, caractérisé en ce qu'il comporte les étapes
5 suivantes :

- on sélectionne un échantillon de documents représentatifs de chaque thème ;

- on identifie, dans les documents sélectionnés, des
10 éléments caractéristiques de chaque thème ;

- on affecte, à chaque élément identifié, un coefficient (R) représentatif de la pertinence de cet élément vis à vis du thème correspondant ;

- pour chaque document (50) à classifier, on
15 identifie lesdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, on calcule, à partir du coefficient affecté à ces éléments, la valeur d'une caractéristique représentative de la pertinence du thème pour ce document
20 (50), pour décider si ce document porte ou non sur ce thème, lesdites étapes d'identification et de calcul étant réalisées automatiquement pour chaque document récupéré sur un réseau informatique ;

- on classe les documents récupérés en fonction des
25 thèmes qui y sont abordés ; et

- l'on stocke les documents classés par thèmes dans des bases de données interrogeables à partir de thèmes
contenus dans une requête ;

et en ce que l'étape d'affectation dudit coefficient
30 à chaque élément identifié comprend les étapes suivantes, pour chaque thème :

- calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème,

- calcul de la fréquence de l'élément dans les
35 documents sélectionné ne portant pas sur ce thème, et

- calcul du rapport entre les fréquences calculées.

2. Procédé selon la revendication 1, caractérisé en ce qu'il comporte en outre une étape de tri des thèmes

selon une arborescence de thèmes et par ordre décroissant des coefficients.

3. Procédé selon la revendication 1 ou 2, caractérisé en ce que l'étape de calcul de la
5 caractéristique représentative de la pertinence du thème d'un document à classifier comprend les étapes suivantes, pour chaque thème :

- on lit la valeur du rapport (R) desdites
fréquences de chaque élément représentatif du thème
10 extrait du document,

- on multiplie les valeurs lues, et

- on affecte le résultat de cette multiplication à la valeur de ladite caractéristique.

4. Procédé selon l'une quelconque des revendications
15 1 à 3, caractérisé en ce que l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil.

5. Procédé selon la revendication 4, caractérisé en
20 ce que la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

$$\text{score} - \text{seuil}_{\text{thème}} = (R_{\text{moy}})n_{\text{thème}}$$

dans laquelle :

25 score - $\text{seuil}_{\text{thème}}$ désigne la valeur de seuil

R_{moy} représente la valeur moyenne des rapports de
fréquences R des éléments du thème et,

$n_{\text{thème}}$ désigne un nombre prédéterminé.

6. Procédé selon la revendication 4, caractérisé en
30 ce que la valeur de seuil est réglée manuellement.

7. Procédé selon l'une quelconque des revendications
1 à 6, caractérisé en ce que les étapes d'identification
des éléments caractéristiques de chaque thème contenu
dans un document (50) sont réalisées au moyen d'une table
35 de hachage.

8. Procédé selon l'une quelconque des revendications
1 à 7, caractérisé en ce que l'on calcule, pour chaque
élément de vocabulaire d'une requête formulée par

l'utilisateur, des coefficients caractéristiques de l'élément par rapport à chaque thème connu et l'on associe à chaque élément les coefficients et les thèmes correspondant, de sorte que lesdits coefficients
5 atteignent une valeur minimale.

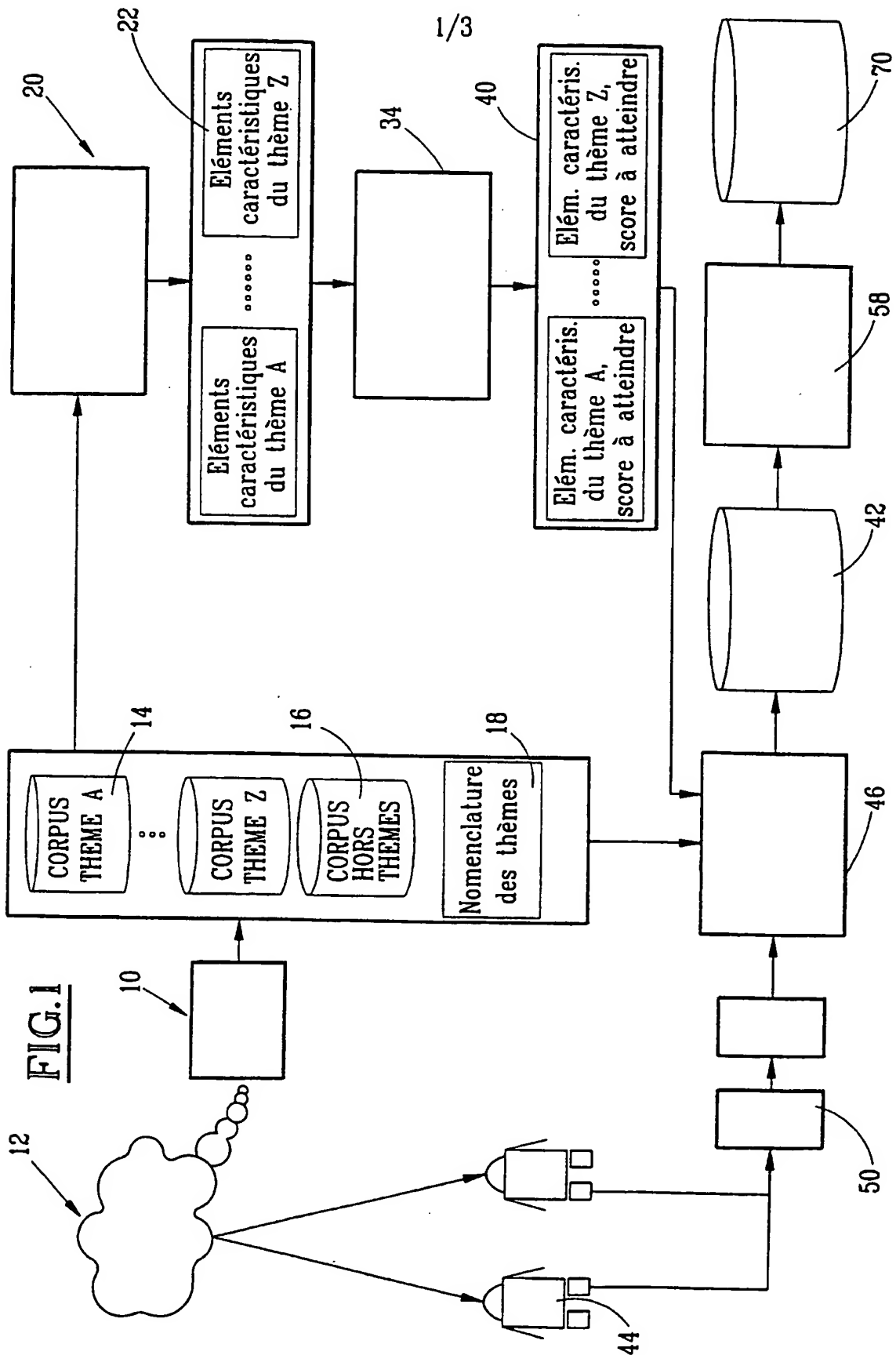
9. Module de classification thématique de documents (50), notamment pour moteur de recherche, caractérisé en ce qu'il comporte une unité centrale de traitement comprenant des moyens de comparaison d'éléments extraits
10 de chaque document avec des éléments caractéristiques de différents thèmes, affectés chacun d'un coefficient (R) représentatif de la pertinence de cet élément pour un thème correspondant, et des moyens de calcul de la valeur d'au moins une caractéristique représentative de la
15 pertinence d'un thème pour ce document, à partir des coefficients desdits éléments caractéristiques qu'il contient, pour décider si ce document (50) porte ou non sur ce thème, ladite unité centrale étant raccordée à des moyens de stockage de documents classés par thèmes,
20 interrogeables à partir de thèmes contenus dans une requête, et en ce qu'il comporte des moyens de calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème, des moyens de calcul de la fréquence de l'élément dans les documents sélectionnés ne
25 portant pas sur ce thème, et des moyens de calcul du rapport entre les fréquences calculées.

10. Utilisation d'un module de classification thématique de documents selon la revendication 9 pour la détermination de thèmes contenus dans une requête
30 formulée par un utilisateur.

11. Utilisation d'un module de classification thématique de documents selon la revendication 9 pour la détermination de thèmes contenus dans des pages récupérées sur un réseau informatique ou dans une requête
35 formulée par un utilisateur et le filtrage des documents récupérés pour interdire la consultation de pages portant sur un ou des thèmes prédéterminés.

12. Utilisation d'un module de classification thématique de documents selon la revendication 9 pour la détermination de thèmes contenus dans une requête formulée par un utilisateur et l'élaboration de profils d'utilisateurs à partir des thèmes sur lesquels porte la requête.

13. Moteur de recherche de documents sur un réseau informatique, comprenant un module d'indexation pour la création et la mise à jour de bases de données thématiques, à partir de documents récupérés sur le réseau informatique, et un module d'interrogation des bases de données thématiques adaptées pour fournir des références de documents correspondant à une requête reçue en entrée, caractérisé en ce qu'il comporte en outre un module de classification thématique selon la revendication 9, associé au module d'indexation.



2/3

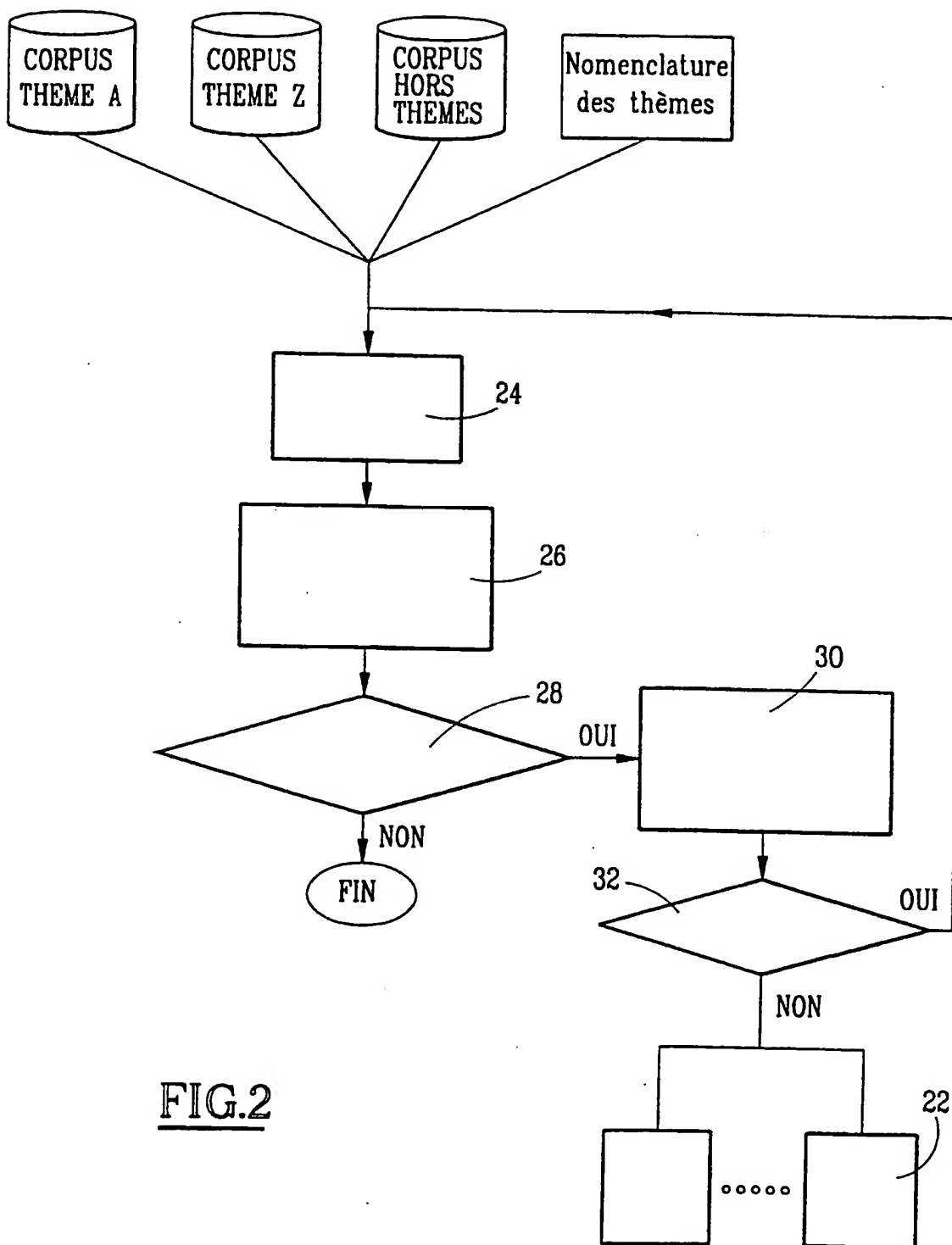
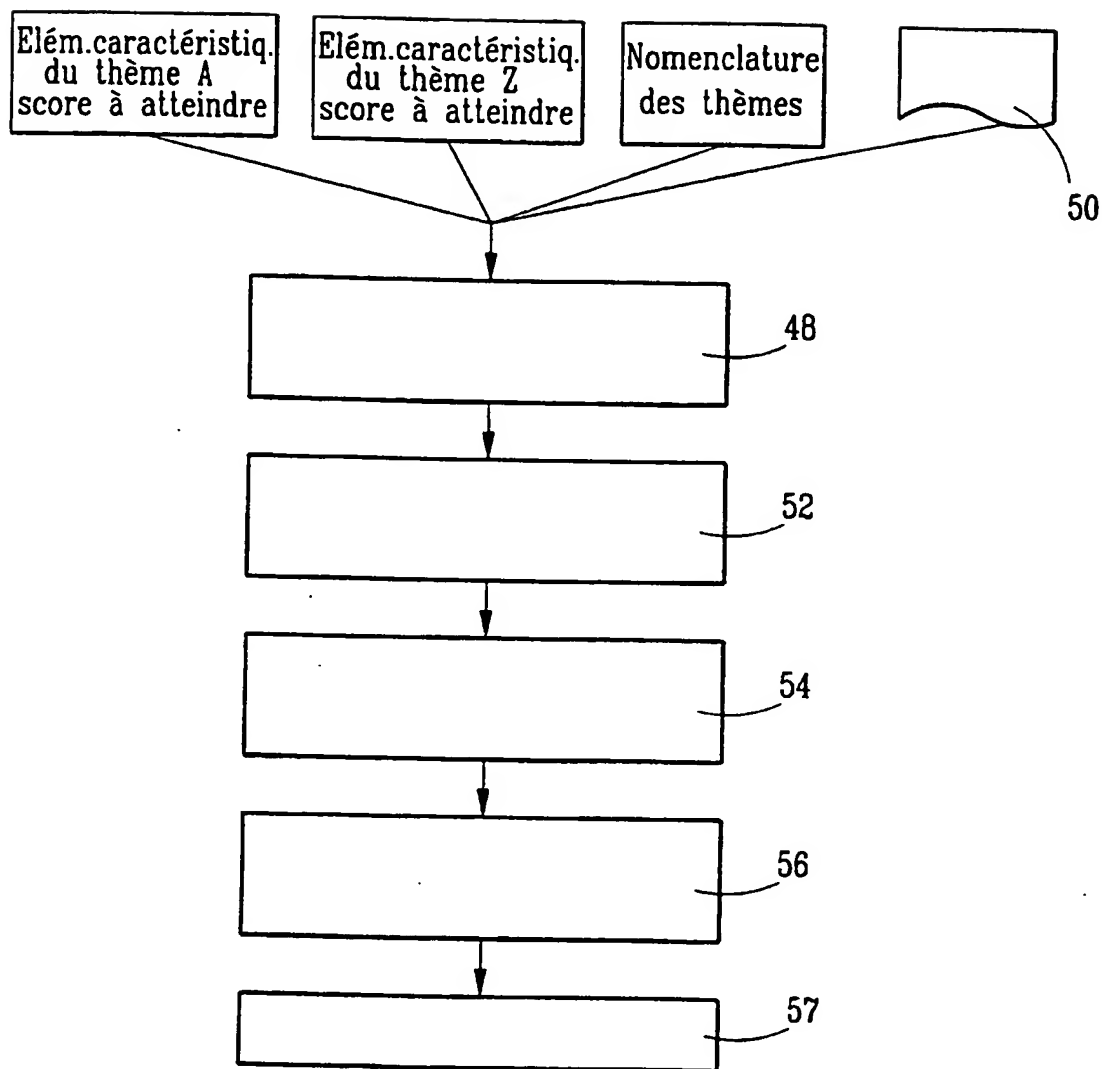


FIG.2

3/3

FIG.3

INTERNATIONAL SEARCH REPORT

International Application No

PCT/FR 00/02640

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

WPI Data, EPO-Internal, PAJ, INSPEC, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 625 767 A (BARTELL BRIAN ET AL) 29 April 1997 (1997-04-29) column 2, line 11 - line 65 column 5, line 15 -column 6, line 31; figures 4,5 column 8, line 54 -column 9, line 28; figure 9	1,8-11
A	CHEKURI C ET AL: "Web search using automatic classification" SIXTH INTERNATIONAL WORLD WIDE WEB CONFERENCE, SANTA CLARA, CALIFORNIA, 'Online! 7 - 11 April 1997, XP002141634 Retrieved from the Internet: <URL:http://www.scope.gmd.de/info/www6/pos ters/725/web_search.html> 'retrieved on 2000-06-30! the whole document --- -/-	1,2,9,13

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

A document member of the same patent family

Date of the actual completion of the international search

15 November 2000

Date of mailing of the international search report

22/11/2000

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Polzer, A

INTERNATIONAL SEARCH REPORT

International Application No
PCT/FR 00/02640

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>CHAKRABARTI S ET AL: "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies"</p> <p>VLDB JOURNAL, AUG. 1998, SPRINGER-VERLAG, GERMANY, vol. 7, no. 3, pages 163-178, XP002141635 ISSN: 1066-8888 abstract page 164, right-hand column, line 30 -page 165, left-hand column, line 24 page 166, right-hand column, line 6 - line 15 page 168, right-hand column, line 6 -page 171, left-hand column, line 40; figures 2-5 page 172, right-hand column, line 15 - line 48</p>	1,7,9, 12,13
A	<p>WO 97 38382 A (BLOCK HANS ULRICH ;SIEMENS AG (DE); BRUECKNER THOMAS (DE)) 16 October 1997 (1997-10-16) abstract page 3, line 31 -page 4, line 27 page 5, line 1 -page 7, line 13; figures 1-4</p>	1,4,9
A	<p>EP 0 822 503 A (MATSUSHITA ELECTRIC IND CO LTD) 4 February 1998 (1998-02-04) abstract page 2, line 28 - line 52 page 4, line 23 -page 5, line 16; figure 1</p>	8,10,11

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/FR 00/02640

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 5625767	A	29-04-1997	AU 5251196 A	02-10-1996
			WO 9628787 A	19-09-1996
WO 9738382	A	16-10-1997	DE 59701176 D	06-04-2000
			EP 0891593 A	20-01-1999
			JP 2000509173 T	18-07-2000
EP 0822503	A	04-02-1998	JP 10049543 A	20-02-1998

RAPPORT DE RECHERCHE INTERNATIONALE

demande internationale No

PCT/FR 00/02640

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
CIB 7 G06F17/30

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)

CIB 7 G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si réalisable, termes de recherche utilisés)

WPI Data, EPO-Internal, PAJ, INSPEC, IBM-TDB

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie *	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	US 5 625 767 A (BARTELL BRIAN ET AL) 29 avril 1997 (1997-04-29) colonne 2, ligne 11 - ligne 65 colonne 5, ligne 15 - colonne 6, ligne 31; figures 4,5 colonne 8, ligne 54 - colonne 9, ligne 28; figure 9	1,8-11
A	CHEKURI C ET AL: "Web search using automatic classification" SIXTH INTERNATIONAL WORLD WIDE WEB CONFERENCE, SANTA CLARA, CALIFORNIA, 'en ligne! 7 - 11 avril 1997, XP002141634 Extrait de l'Internet: <URL:http://www.scope.gmd.de/info/www6/pos ters/725/web_search.html> 'extrait le 2000-06-30! le document en entier	1,2,9,13
	-/-	



Voir la suite du cadre C pour la fin de la liste des documents



Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

- "A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent
- "E" document antérieur, mais publié à la date de dépôt international ou après cette date
- "L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)
- "O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens
- "P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée

15 novembre 2000

Date d'expédition du présent rapport de recherche internationale

22/11/2000

Nom et adresse postale de l'administration chargée de la recherche internationale
Office Européen des Brevets, P.B. 5818 Patentaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3018

Fonctionnaire autorisé

Polzer, A

C.(suite) DOCUMENTS CONSIDERES COMME PERTINENTS		
Catégorie	Identification des documents citée, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	<p>CHAKRABARTI S ET AL: "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies" VLDB JOURNAL, AUG. 1998, SPRINGER-VERLAG, GERMANY, vol. 7, no. 3, pages 163-178, XP002141635 ISSN: 1066-8888 abrégé page 164, colonne de droite, ligne 30 -page 165, colonne de gauche, ligne 24 page 166, colonne de droite, ligne 6 - ligne 15 page 168, colonne de droite, ligne 6 -page 171, colonne de gauche, ligne 40; figures 2-5 page 172, colonne de droite, ligne 15 - ligne 48</p>	1,7,9, 12,13
A	<p>WO 97 38382 A (BLOCK HANS ULRICH ;SIEMENS AG (DE); BRUECKNER THOMAS (DE)) 16 octobre 1997 (1997-10-16) abrégé page 3, ligne 31 -page 4, ligne 27 page 5, ligne 1 -page 7, ligne 13; figures 1-4</p>	1,4,9
A	<p>EP 0 822 503 A (MATSUSHITA ELECTRIC IND CO LTD) 4 février 1998 (1998-02-04) abrégé page 2, ligne 28 - ligne 52 page 4, ligne 23 -page 5, ligne 16; figure 1</p>	8,10,11

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Requête internationale No

PCT/FR 00/02640

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 5625767 A	29-04-1997	AU 5251196 A WO 9628787 A	02-10-1996 19-09-1996
WO 9738382 A	16-10-1997	DE 59701176 D EP 0891593 A JP 2000509173 T	06-04-2000 20-01-1999 18-07-2000
EP 0822503 A	04-02-1998	JP 10049543 A	20-02-1998